Predicting Student Performance in Object Oriented Programming Using Decision Tree : A Case at Kolej Poly-Tech Mara, Kuantan

Mohd Hanis Rani[1*], Abdullah Embong[1],

[1] Faculty of Computer System and Software Engineering,
Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300 Kuantan Pahang, Malaysia
bulat_is@yahoo.com , abdullahbe@ump.edu.my

Abstract. The paper focuses on prediction of student learning performance in object oriented programming course using data mining technique based on a dataset obtained from Kolej Poly-Tech Mara (KPTM), Kuantan. The objective was to identify and implement the most accurate algorithm for the KPTM dataset and to come up with a good prediction model using decision tree technique. The most relevant rules were identified from the model. The dataset was run through some pre-processing such as data cleaning, data reduction and discretization. The experiments were conducted using machine learning software Weka 3.6.9. The first experiment was to test the clean dataset with seven classification techniques. Accuracy plays an important role to prove the best classification technique by using correctly classified instance as an indicator. Using 10-fold cross validation for each algorithm, it was found that decision tree was the best algorithm with 83.6944% correctness. The second experiment was conducted to find the best model among the percentage split where the best percentage split produced the best model accuracy. The experiment with 50% of data training and 50% of data testing in percentage split produced higher accuracy where the percentage of correctly classified instance was 76.2557%. The rules were extracted from the model and after the analyses were conducted the result showed that the domain factors of student performance were class attendance and the performance of the previous semester.

1 Introduction

Even in the cyber age, education still plays very important role in the development and modernization of the country. Education leads to sustainable quality graduates capable of providing a quality workforce for the country. In computer science, the quality of learning has grown in tandem with technological growth especially in the use of programming. Object-oriented programming (OOP) is one of the core courses in computer science and technology, which is also one of the most important specialty courses for science and engineering university students [5]. At Kolej Poly-Tech Mara (KPTM), OOP is a major subject for the students in the Diploma of Information

Technology programme. The problem is many students failed or did not perform well in this subject.

There are many factors which contribute to the student failure such as the student lack of understanding, absenteeism from class and the student weak education background. One of the ways to improve the student performance is for the instructors to identify the group of students who might not perform well at the early stage of learning. From there the instructor can focus on the group in order to help them to improve their performance. Thus, in this case making the prediction of student learning performance is a major step in identifying the potential group that needs further help such as extra classes or special tutorials and assignments.

2 Performance Prediction

Usually the lecturer can predict, to a certain degree, the future performance levels of students based on their performance in Mathematics and English at SPM (Malaysian equivalent of O-Level), soft skill such as attendance and a few other attributes. Indirectly, advice and suggestions can be given to poor students.

Data mining is a step in the knowledge discovery from database process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data [1]. Data mining is defined as a logical process that is used to search through large amount of data in order to find the useful patterns that were previously unknown. The useful patterns that are found will represent the new knowledge [3]. Most of data mining methods are based on tried and tested techniques from machine learning, pattern recognition and statistics such as classification, clustering, and regression [2]. Data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data [4],[5]. According to Han et al., [11], the process of finding the pattern in the data set is done by using data mining techniques.

Classification is a part of data mining [14],[15]. Classification involved the process to analyze the pattern of data in training set to find out an accurate model. The knowledge analysis from the results will be evaluated to generate a new model. Classification and prediction are related techniques [13]. Classification has many algorithms such as back propagation, association-based classification, decision tree, Bayesian classification and rough set theory but the most popular classification method is decision tree and Bayesian classification [12]. Decision tree structures are a common way to organize classification schemes[11].

There have been a few studies done in constructing prediction models in education for various purposes. Behrouz et al., [6] use classification methods and techniques that are available in data mining in order to predict the performance of students at Michigan State University (MSU). He used various classification techniques such as Multilayer Perceptron, Quadratic Bayesian, Parzen-windows, I-nearest neighbor (I-NN), Decision and Tree K-nearest neighbor (k-NN). This study combines a number of classifications and the results of tests performed provide a significant improvement in the measurement of the level of classification. Additionally, learning the characteristics of classification using genetic algorithm has

improved the accuracy of predictions made. The study was conducted on 227 students' data and this study is the prediction of student performance based on assigned homework.

Chun-Teck et al., [7] conduct the prediction on pre-university students on mathematics achievement. The study used three methods, which are the Generalized Regression Neural Network (GRNN), Classification, Back-propagation Neural Network (BPNN) and Regression Tree (CART) in order to predict the students' mathematics achievements. The study consists of two parts, *i.e* to predict the students' mid-semester assessment result and the final examination result. The output based on models' accuracy is evaluated to identify the best model. The findings reveal that BPNN outperforms other models with an accuracy of 66.67% and 71.11% in predicting the mid-semester evaluation result and the final examination result respectively. The studies used 180 students' data who enrolled in the foundation of engineering at Multimedia University.

Arshad et al., [8] conducted a study to predict the engineering students performance at the University of Engineering and Technology, Peshawar using 203 students' data. The association between the predictors which is entry test scores and overall merit and the criterion such as academic achievements or scores of engineering students from first to final year were analyzed using appropriate statistical procedure. The findings indicate that there is significant relationship between entry test scores and overall merit with the academic achievement of engineering students. Umeh et al., [9] have succeeded in identifying the characteristics of weak students by conducting a survey using Bayesian classification techniques with 600 students' data. Shaeela et al., [10] worked on data mining model for higher education system to make predictions about the classroom performance in relation to students' attendance. All those finding show that data mining can be used to predict student performance.

3 Methodology

Raw data of 4405 students who enrolled in the Diploma of Information Technology, KPTM was collected from the academic department. The data involve 28 attributes, among others are 'age', 'gender', 'country', selected SPM results and previous semester results. The data also include 'attendance' as a soft skill to ensure the accuracy of the prediction. The raw data was run through data preprocessing such as data cleaning, data reduction, and discretization to ensure the quality of the mining results.

According to Han et al., [11], dirty data can be caused by many issues such as the problems arising from human, IT hardware and software failure, data entry errors in the system, data transmission errors and mistakes that are not relevant to the data collection. There are a few solutions to the dirty data problem, one of them is value replacement using modes [11]. The mode refers to the list of number that occurs most frequently in the dataset. Data reduction is a process of removing the unused attributes from the data set. Data reduction can enhance the effectiveness of data

mining and modeling. Table 2 shows the list of attributes left after the data reduction exercise.

Table 2. List of the attributes.

Data description		
No.	Variable	Description
1	Jantina	Male or female
2	Negeri	State in Malaysia
3	BM	SPM Malay Language
4	BI	SPM English Language
5	MAT	SPM Mathematic
6	MATTAM	SPM Additional Mathematic
7	SEJ	SPM History
8	AGAMAMORAL	SPM Islamic Religion / Moral
9	TMK 121	Personal Computer Technology Subject
10	TMA 111	Introduction to Programming Subject
11	TMA 222	Object Oriented Programming Subject
12	STATUS	Status for semester 1
13	CGPA	Grade for semester 1
14	KEHADIRAN	Status of attendance

Data discretization is a process of dividing the range of continuous attributes into intervals to reduce the data size. It helps to prepare the analysis in the prediction.
The CGPA attribute will be converted to the category that is easier to understand for the purpose of discretization process. All the data input will be represented in the form of specific categories to facilitate data mining.

4 Implementation

The experiment was conducted using Weka (Waikato Environment for Knowledge Analysis) software version 3.6.9, developed by University of Waikato, New Zealand.
The first experiment was to test the clean dataset with seven classification techniques i.e. Naïve Bayes, Logistic, Decision Table, Classification Via Clustering, OneR, User Classification and Decision Tree. Accuracy plays an important role to prove the best classification technique by using correctly classified instance as an indicator. Correctly classified instance shows the percentage of data that was correctly classified by the algorithm. Higher percentage of correctly classified instance mean the model has a higher accuracy.

The clean dataset were processed using 10-fold cross validation. All values of correctly classified instance were compared to determine the best technique to be selected. The technique with the highest percentage of correctly classified instance will be selected as the technique in the development of the model. Table 3.0 shows the result of correctly classified instance through seven classification techniques. From the table it is clear that the highest percentage of correctly classified instance

was obtained by decision tree. The results prove that, compared to the other classification techniques, the decision tree is the best classification technique.

Table 3. Result of correctly classified instance through seven classification techniques.

Classification Technique	Correctly classified instance
Naïve Bayes	64.7662
Logistic	65.9065
Classification Via Clustering	41.6192
Decision Table	81.87
One R	64.4242
User Classification	44.9259
Decision Tree	83.6944

The second experiment was conducted to find the best model among the percentage split where the best percentage split produced the best model accuracy. Based on Han et al., [11], accuracy can be estimated using one or more test sets that are independent of the training set where estimation techniques that can be used such as cross-validation and percentage split. The most accurate model will produce the best and strong rules. To run the experiment, the model development was divided into several percentages split and each percentage split will define a model. The model with the highest percentage of correctly classified instance will be chosen as the best model and the percentage split will be observed. Table 4.0 shows the result of the model through percentage split.

Table 4 . Result of the model through percentage split.

Model	Percentage Split %		Correctly Classified Instance
	Training	Testing	
A	90	10	76.1364
B	80	20	74.8571
C	70	30	73.7643
D	60	40	73.2194
E	50	50	76.2557
F	40	60	72.4335
G	30	70	69.3811
H	20	80	67.2365
I	10	90	63.6248

Model E was chosen as it gave the highest value of correctly classified instance. It uses 50% of data training and 50% of data testing in percentage split. The next process is to extract the rules from the model. From the chosen model, the tree will be observed in details to select the most relevant rules which will become the output in this research. There exists a technique to extract the right rules. A rule is created for each path from the root to the leaf of the tree and each attribute-value pair along a path forms a conjunction. The leaf node holds the class prediction.

Base on the rules that were extracted from the model, new knowledge such as domain factor and the attribute behavior were produced after some analysis was done for each attribute. The analysis was conducted based on the rules to see how the attributes play a role in producing tips. Fig.1 illustrates the analysis from the rules. The attribute with the highest value would contribute the most to the rules, where in this case it is the student attendance. The second highest is “TMA111”, followed by “TMK121”, “BI”, "MAT", “CGPA" and "MATTAM".

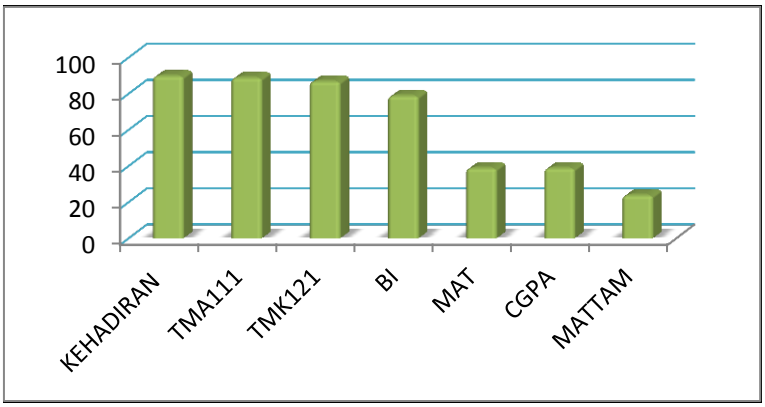

Fig. 1 . The value of each attribute that was involved in classification rules.

5 Result and Discussion

In the context of the rules obtained, an interesting set of rules covered attributes regarding the subjects taken by students for the first semester, several SPM subjects like Mathematic, Additional Mathematic, English and attendance.

Attribute “kehadiran” or class attendance become the root of the decision tree model based on the training dataset and that made attendance as the highest contributing attribute. Attendance gave a big impact on the rules such as *if attendance is good then **t**he result for the object oriented would be excellent*. This shows the domain factor of the class attendance as it influences the result of the OOP subject at KPTM Kuantan.

The attribute “TMA111” (Introduction to Programming) and “TMK121” (Personal Computer Technology) represent the subject of the previous semester which all students have to enroll. Both attributes are the second level attributes which may influence the rules. “TMA111” syllabus contains the introduction to programming where the student can get the basic knowledge in programming and understand the programming subject in advance. It will help the student to gain more understanding of OOP. Same goes to “TMK121” where the syllabus of the subject contains the structure of programming fundamental. If the student have a problem with the attendance but got good result in both subjects, then the student may pass the OOP subject as stated in therule:, *if kehadiran = "teruk" and TMA111="lulus" and TMK121 ="kepujian" then TMA222 = "lulus"*.

English subject also effected the rules since English influences the student learning performance in OOP. This makes sense because students have to study OOP in English. Some students could not understand the learning because of the language barrier. Besides the lectures, all the notes and reference materials were provided in English. Meanwhile Mathematic, Additional Mathematic and previous semester results were found not to play important role in the prediction of the students’ performance in OOP.

6 Conclusion

A study has been conducted on prediction of student learning performance in OOP course using data mining technique based on a dataset obtained from KPTM, Kuantan. The objectives were to identify and to implement the most accurate algorithm for the KPTM dataset and to come up with a good prediction model using decision tree technique. The most relevant rules have been identified from the model. Accuracy plays an important role to prove the best classification technique and it was found that decision tree was the best algorithm with 83.6944% correctness. The best model among the percentage split were 50% of data training and 50% of data testing that produced higher accuracy where the percentage of correctly classified instance was 76.2557%.

The rules were extracted from the model and after the analyses were conducted the result showed that the domain factor of student performance was class attendance and the students’ performance in the previous semester. The factors which contribute to the student failure were absenteeism from class. It is important for the instructors to identify the group of students who might not perform well at the early stage of learning in order to improve their performance by focusing extra help on them

Acknowledgments. We wish to acknowledge the contribution of several individuals who had given their support during the research and ERGS Grant: CSTWay: A Computational Strategy for Sequence Based T-Way Testing for supporting this paper.

References

1. Fayyad, Piatetsky-Shapiro & Smyth : Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KKD-96), Portland, Oregon, August 2-4, (1996) AAAI.
2. Fayyad, Piatetsky-Shapiro & Smyth : From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence (1996) 0738-4602-1996
3. Mrs. Bharati M. Ramageri : Data Mining Techniques And Applications, Indian Journal of Computer Science and Engineering, Volume. 1 No. 4 , (2010) 301-305.
4. S. Mitra, S. K. Pal, P. Mitra : Data Mining in Soft Computing Framework: A Survey, IEEE Transactions On Neural Networks, Vol.13, No. 1. (2002)
5. Jie Anquan, Li Yuqing, Chen Bailiang, Ye Jihua, Zou Jie : The Education Reform and Innovation of Object Oriented Programming Course in Normal, The 5th International Conference on Computer Science & Education Hefei, (2010) 978-1-4244
6. Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer, William F. Punch : Predicting student performance: an application of data mining methods with the educational web-based system lon-capa, 33rd ASEE/IEEE Frontiers in Education Conference, (2003) 0-7803-7444-4
7. Chun-Teck. L., Lik N. N., M. Daud Hassanc, Wei W. G., Check Y. L, Noradzilah I., Predicting Pre-university Students' Mathematics Achievement, International Conference on Mathematics Education Research, (2010) 299–306.
8. Arshad A., Umar A. : Predictability of engineering students' performance at the University of Engineering and Technology, Peshawar from admission test conducted by educational testing and evaluation agency ETEA, NWFP, Pakistan, Procedia Social and Behavioral Sciences 2 (2010) 976–982
9. Umesh Kumar Pandey S. Pal, : Data Mining : A prediction of performer or underperformer using classification, International Journal of Computer Science and Information Technologies, Vol. 2 2., (2011) 686-690.
10. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M.Inayat Khan : Data Mining Model for Higher Education System, European Journal of Scientific Research, ISSN 1450-216, Vol.43 No.1, (2010) pp.24-29
11. Han. J and Kamber M. : Data Mining: Concepts and Techniques. 2nd ed. San Francisco, California: Morgan Kaufman (2006).
12. Samuel Odei Danso : An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain, Master Degree Thesis, Bournemouth University (2006)
13. C. Romero, S. Ventura : Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications 33, 135–146 (2007)
14. C. Romero, S. Ventura : Educational Data Mining: A Review of the State of the Art, IEEE Transactions On Systems, Man, And Cybernetics, Volume 40, No. 6 (2010)
15. K.Srinivas, B.Kavihta Rani, A. Govrdhan : Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, International Journal on Computer Science and Engineering Vol. 02, No. 02. (2010)